UNLV | UNIVERSITY LIBRARIES

Publications (YM) | Yucca Mountain

9-2003

Final report of Task #5: Current document index system for document retrieval investigation

Thomas Nartker
University of Nevada, Las Vegas, tnartker@cs.unlv.edu

Kazem Taghva
University of Nevada, Las Vegas

Julie Borsack
University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/yucca_mtn_pubs

 Part of the Library and Information Science Commons

Repository Citation

Nartker, T., Taghva, K., Borsack, J. (2003). Final report of Task #5: Current document index system for document retrieval investigation.
Available at: https://digitalscholarship.unlv.edu/yucca_mtn_pubs/38

This Article is protected by copyright and/or related rights. It has been brought to you by Digital Scholarship@UNLV with permission from the rights-holder(s). You are free to use this Article in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you need to obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/or on the work itself.

This Article has been accepted for inclusion in Publications (YM) by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.

Final Report of Task #5: Current Document Index System for Document Retrieval Investigation

1998-2003 DOE Coop

Tom Nartker, Kazem Taghva, and Julie Borsack

September 2003

UNLV
Information Science
Research Institute

EXECUTIVE SUMMARY

In Part I of this report, we describe the work completed during the last fiscal year (October 1, 2002 thru September 30, 2003). The single biggest challenge this past year has been to develop and deliver a new software technology to classify Homeland Security Sensitive documents with high precision. Not only was a satisfactory system developed, an operational version was delivered to CACI in April 2003. The delivered system is called the Homeland Security Classifier (HSC).

In Part II we give an overview of the projects ISRI has completed during the first four years of this cooperative agreement (October 1, 1998 thru September 30, 2002). Each of the deliverables associated with these projects has been thoroughly described in previous reports.

TABLE OF CONTENTS

Final Report of Task #5:
Current Document Index System
for Document Retrieval Investigation

1998-2003 DOE Coop

Tom Nartker, Kazem Taghva, and Julie Borsack
September 2003

Introduction

This report summarizes the work accomplished by UNLV/ISRI for Task #5 of the 1998-2003 DOE Coop. In Part I of this report, we give a detailed description of each specific task accomplished during the last year. We discuss the significance of this work and give a list of the deliverables provided.

In Part II, we give an overview of tasks performed during the previous four years. A detailed description of the studies completed is shown in previous annual and quarterly reports. Appendix A summarizes the deliverables provided during the five year period. Appendix B, C, and D show publications, graduate students, and undergraduate students supported.

PART I. Work completed for FY 02/03

I-1. Specific Tasks for FY02/03

The specific tasks planned for FY02/03 fell into one of six different categories. These categories are shown as 1 through 6 below. Very early in the year, it became clear that tasks associated with category number one were more important and took priority over other tasks. Although several tasks were completed that were associated with categories 2 and 3, no effort was devoted to categories 4, 5 or 6.

1. Develop methodologies for classification of DOE records and e-mail to facilitate LSN loading, search and retrieval, and other related activities.

2. Assist in developing, refining, and implementing all activities associated with the capture of DOE records.

3. Participate in LSN related discussions with the NRC and the DOE as requested.

4. Evaluate the NRC electronic courtroom requirements and determine how the DOE might best meet those requirements.

5. Investigate methods for improving the search and retrievability of records within the DOE record system.

6. Perform an analysis of the DOE records system to determine future use of records and the record formats that should be employed.

I-2. Studies and Results

The most important task undertaken this year has been to develop a technology to support high precision automatic "classification" of DOE records (i.e., category #1 above). In particular, a system was needed to process a stream of documents to filter out those that contained, or might contain, Homeland Security Sensitive information. The important requirement was that NO sensitive documents would be passed through the system.

There are about a half-dozen products available to perform such classification operations but NONE of these can provide output that is 100% free of sensitive information. In addition to providing 100% precision output, the new system needed to include a semi-automated post-processing facility for separating non-sensitive from sensitive documents in the filtered stream to achieve complete identification of sensitive information at minimum cost.

The background research was completed in the Fall of 2002. A first prototype system was constructed and tested in December. The new system passed a major performance test in mid-December and the ISRI staff spent January thru March producing an industrial strength product for DOE use. This system, called the Homeland Security Classifier (HSC) was delivered to and installed on CACI computers in April 2003.

Having produced one system capable of high precision classification, we quickly focused on the problem of separating LSN-relevant messages from non-relevant messages in a stream of e-mail messages. A new system was constructed and tested during the summer of 2003. This system, called the LSN e-mail classifier (LEC) was delivered on October 1, 2003.

Initial research regarding the classification of documents containing Privacy Act information has begun. At this time, ISRI staff awaits arrival of example documents that can be used for training, testing, and validation. No reports regarding Privacy Act work are yet available.

Although the main focus of ISRI work this year has been on classification problems, ISRI staff have also worked on problems relating to "activities associated with the capture of DOE records" (i.e., category 2 above). Most of this activity has been associated with enhancements to the MANICURE post-processing system. MANICURE was adopted more than a year ago for automatic correction of OCR errors in the DOE conversion system.

The improvements to MANICURE include:

- Changes to the adaptive confusion matrix eliminating the problem of certain poor quality (and/or long documents) taking too long to process,
- Changes to utilize the most recent versions of Ispell and the Berkeley DB library,
- Addition of the pre-built Ispell dictionaries for Solaris and Linux, using the MANICURE wordlist, which makes installation easier,
- Installation of Autotag functionality in Ppsys,

- Changes to support switching of PDOC and LDOC files from SGML to XML in the next version, and
- Linux support.

Each of these improvements has been delivered and installed on CACI machines.

Another ISRI task associated with the capture of records, was to establish OCR quality control procedures for the DOE conversion system. The NRC goal for accuracy of OCR output has been set in terms of the accuracy of "non-stopwords" in the output. The NRC criteria is for 95% (or better) correct non-stopwords.

Because MANICURE produces output statistics regarding the number of misspelled and the number of correctly spelled words recognized on a page, it is possible to calculate an estimate of the percentage of correctly spelled non-stopwords in each document processed. With some training and calibration, ISRI staff have established a "threshold" on this estimate that insures 95% correct non-stopword accuracy for all documents processed. This QC procedure has been implemented and is another of the enhancements to MANICURE this year.

Finally, ISRI staff have also "participated in LSN related discussions with the NRC" (category 3 above). In December, Kazem Taghva, Julie Borsack, and Tom Nartker participated in the EIE Technical Exchange Meeting with NRC staff. As part of David Warriner's presentation, Kazem gave a description of the Homeland Security Classifier. He concluded by indicating the performance obtained on preliminary tests. In addition, ISRI staff have participated in weekly conference calls of the NRC Working Group.

I-3. Significance

The significance of the reports and systems produced by ISRI staff over the five year period of this contract can be measured in three different areas.

1. A reduction in the cost of providing records to the NRC for use in the LSN,

2. An increase in the quality of the records delivered, and

3. The availability of information regarding document conversion and document retrievability to support
   effective management decision making within the OCRWM program.

I-4. Deliverables

Appendix A shows a list of all deliverable reports and systems provided during the five years of the Coop Agreement.

I-5. Publications and Presentations

During FY 02/03, several ISRI authored journal articles and conference papers were published. A list of these is shown below. Appendix B shows a list of publications over the five years of this contract.

Tom Nartker, Kazem Taghva, Ron Young, Julie Borsack, and Allen Condit, OCR Correction based on Document Level Knowledge, Proceeding of SPIE 2003, January 2003, pages103-110.

Kazem Taghva and Jeff Coombs, Do Thesauri Enhance Rule-Based Categorization for OCR Text?, Proceeding of SPIE 2003, January 2003, pages 111-119.

Kazem Taghva, Julie Borsack, Jeffrey Coombs, Allen Condit, Steve Lumos, and Tom Nartker, Ontology-based Classification of Email, Proc. Intl. Conf. on Information Technology: Computers and Communications, pages 194-198 Las Vegas, NV, April 2003.

PART II. Overview of Activities from 10/01/98 thru 9/30/02

II-1. Specific Tasks for FY's 98/99, 99/00, 00/01, and 01/02

The tasks undertaken by ISRI over the first four years of the Coop program fall into three main areas. These are:

1. Studies of existing information technologies related to the LSN,
2. Studies of DOE information systems, and
3. Studies related to the LSN and to interactions with the NRC.

II-2. Studies and Results

The most extensive technology studies have been associated with Information Retrieval (or search) systems. ISRI has not only conducted an in depth performance comparison of existing commercial search systems (January 2000), but it has also investigated many other search related issues. Examples include thesauri-aided retrieval (May 1998), usefulness of manually assigned keywords (December 2000), and retrievability from manually zoned vs. automatically zoned collections (January 2002).

A second technology studied thoroughly was Optical Character Recognition. Although ISRI's most comprehensive studies of OCR systems predate this Coop agreement, a study of the OCR accuracy produced by the DOE conversion system (May 2002) is an example of OCR based studies. Our knowledge of the performance of commercial OCR systems has been valuable as part of many other studies.

The newest technologies studied have been electronic information classifiers. The use of Autonomy to assign RIS documents into OPRRS categories (May 2000) and the effectiveness of Autonomy to separate inclusionary from exclusionary documents (July 2001) are good examples. In fact, a thorough comparison of several commercial classifiers was also conducted but the results were never published because of agreements with the vendors. The main conclusion derived from this study was that no existing classification system was capable of producing even one output stream that was100% precise.

The main DOE system studied was the system to convert paper documents into electronic form (May and June 2002). These studies, and the search studies mentioned above, provided the background necessary to convince NRC staff to focus their goal for acceptable accuracy on "non-stopword" accuracy and not on "character" accuracy.

Another DOE system studied was the Lotus Notes email system used within YMP. A study of elimination of duplicate messages from archived electronic mail (October 2000) is an example.

Finally, there have been studies directly related to the LSN or to interactions with the NRC. Although there are many document delivery issues involved, studies involving semi-automatic quality control of OCR output is the most recent example (January 2003).

II-3. Significance

Perhaps the most profound value resulting from the summation of these studies has been in giving ISRI staff the background needed to develop the high precision classifiers described in Part I. The short term economic value of the two systems delivered during the last year of the Coop is significant.

An equally important outcome, although less visible, is the management information needed by YMP staff to make informed decisions regarding the efficiency and quality of information systems needed to prepare documents for inclusion in the LSN.

II-4. Deliverables

Appendix A shows a list of all deliverable reports and systems provided during the five years of the Coop Agreement.

II-5. Publications and Presentations

Appendix B shows a list of ISRI authored journal articles and conference papers published during the five years of the Coop Agreement.

Appendix A.

SUMMARY OF DELIVERABLE REPORTS FROM 1998 THRU 2003

SUMMARY OF DELIVERABLE REPORTS AND SYSTEMS FROM 1998 THRU 2003

REPORTS

The Effectiveness of Thesauri-Aided Retrieval using BASISplus, the LSS Thesaurus, and the LSS Prototype Document Database, May 1998

AN Evaluation of Commercial Speckle and Skew Removal Packages, October 1998

ISRI Activities in Support of YMSCO Licensing and Records: Full-Text IR/DBMS Evaluation

A Comparison of LiveLink8 with DOE/LSN Requirements, October 1999

A Comparison of InQuery 5.1 with DOE/LSN Requirements, October 1999

A Comparison of RetrievalWare 6.6.2 with DOE/LSN Requirements, November 1999

A Comparison of Fulcrum SearchServer 3.7e with DOE/LSN Requirements, November 1999

A Comparison of Thunderstone Texis 2.6.930169407 with DOE/LSN Requirements, December 1999

A Comparison of Basis 8.3 with DOE/LSN Requirements, December 1999

IR SYSTEM EVALUATION: A Comparative Report, January 2000

On the Use of Autonomy to Assign RIS Documents into OPRRS Categories, May 2000

Final Report on the Elimination of Duplicate Messages in Archived Electronic Mail, October 2000

Determining the Usefulness of Manually Assigned Keywords for a Vector Space System, December 2000

Analysis of the Effectiveness of Autonomy to Separate Inclusionary and Exclusionary Documents, July 2001

Regulatory Guide 3.69 Topical Guidelines: Clarifications, Considerations, and Comments, September 2001

Proposed Plan for Investigating the Retrivability from Manually Zoned vs. Automatically Zoned Collections, January 2002.

OCR Accuracy Produced by the Current DOE Document Conversion System, May 2002

Retrieval of Documents from the Current DOE Document Conversion System, May 2002.

Evaluation of the Current DOE Document Conversion System: A Study of Retrievability, June 2002.

Measuring and Delivering 95% Non-Stopword Document Accuracy, August 2003

SYSTEMS

The MANICURE post-processing system.
A sequence of versions of this system have been delivered. The most recent version includes metrics to automate QC of OCR output.

The Homeland Security Classifier (HSC)
Verson 1

The LSN E-mail Classifier (LEC)
Version 1

Appendix B.

ISRI PUBLICATIONS

ISRI PUBLICATIONS

Taghva, K, (with Borsack, J., and Condit A.), Autotag: A Tool for Creating Structured Document Collection from Printed Materials, Electronic Publishing '98, Lecture Notes in Computer Science 1375, Springer-Verlag, pp. 420-431.

Taghva, K., Borsack, J., and Condit A. MANICURE Document Processing System, Proceedings of SPIE Symposium on Electronic Imaging 1998, pp. 179-184.

Taghva, K, (with Gilbreth J.), Finding Acronyms and Their Definitions, Int. Journal of Document Analysis and Recognition, vol. 1, NO. 4, pp. 191-198 (May 1999).

Taghva, K., Borsack, J., and Condit A. Effectiveness of Thesaurus-Aided Retrieval, Proceedings of SPIE Symposium on Electronic Imaging, vol. 3651, pp.134-140. (Jan. 1999).

Taghva, K. Tom Nartker, Julie Borsack, Allen Condit. ISRI Test Collection, Proceedings of SPIE Symposium on Electronic Imaging, 2000 pp. 157-164.

Taghva, K. Tom Nartker, Julie Borsack, Steve Lumos, Allen Condit, and Ron Young. Evaluating Text Categorization in the Presence of OCR Errors, Proceedings of SPIE Symposium on Electronic Imaging, 2001, pp. 68-74.

Taghva, K. and Jeff Coombs, Hairetes: A Search Engine for OCR Documents, Lecture Notes in Computer Science 2423, Springer Verlag, pp. 412-422.

Kazem Taghva and Eric Stofsky OCRSpell: An Interactive Spelling Correction System for OCR Errors Int. Journal on Document Analysis and Recognition, March 2001, pages125-137.

Kazem Taghva, Thomas Nartker, Julie Borsack, and Allen Condit, Determining the Usefulness of Manually Assigned Keywords for a Vector Space System, Proc. of the IEEE Intl. Conf. on Information Technology: Coding and Computing, April 2002, pages 242-246.

Kazem Taghva, Julie Borsack, Tom Nartker, Steve Lumos, and Allen Condit, Managing Occupational Medicine Historical Data, Proceedings of METMBS `02, pp. 256-260.

Tom Nartker, Kazem Taghva, Ron Young, Julie Borsack, and Allen Condit, OCR Correction based on Document Level Knowledge, Proceeding of SPIE 2003, January 2003, pages103-110.

Kazem Taghva, and Jeff Coombs, Do Thesauri Enhance Rule-Based Categorization for OCR Text?, Proceeding of SPIE 2003, January 2003, pages111-119.

Kazem Taghva and Julie Borsack and Jeffrey Coombs and Allen Condit and Steve Lumos and Tom Nartker, Ontology-based Classification of Email, Proc. Intl. Conf. on Information Technology: Computers and Communications, pages 194-198 Las Vegas, NV, April 2003.

Appendix C.

GRADUATE STUDENTS SUPPORTED & THESIS TOPICS

.

GRADUATE STUDENTS SUPPORTED & THESIS TOPICS

Elena Dimitrova	M.S. Computer Science, Thesaurus-aided retrieval	1998
Lydia Macowski	M.S. Computer Science, Text Categorization of Noisy Data	2000
Jeff Coombs	M.S. Computer Science, Rule Based Categorization	2002
Tong Feng	M.S. Computer Science, Large Scale DB Modeling: Extended ER and UML	2002
Jun Liu	M.S. Computer Science, Interactive Ground Truth tools	2002
Zumreta Maslesa	M.S. Computer Science, Large Scale DB Modeling: Developing XML Schema	2002
Min Xu	M.S. Computer Science, Content Based Image Retrieval	2002
Ying Yang	M.S. Computer Science, DB Modeling: Discovering Entities and Relationships	2002
Zhenxing Mao	M.S. Computer Science, Relational Detection and Correction of OCR Errors	2003
Russ Buckley	M.S. Computer Science, Stemmers for Cross Language Retrieval	2003
Renato Martleto	M.S. Computer Science, Rule Processing Using Ontology	2003
Qin Liu	M.S. Computer Science, Detection & Removal of Headers and Footers	2003

Appendix D.

UNDERGRADUATE STUDENTS SUPPORTED & MAJOR

UNDERGRADUATE STUDENTS SUPPORTED & MAJOR

Alma Hanson	computer science
Anna Aurelio	business
Benjamin Minx	geology
Brian Earley	geology
Colleen Smith	biology
David Shields	geology
Elizabeth Earle	computer science
James Ford	geology
Katrina Rodriguez	international relations
Kiersten Maholtz	geology
Kim Caprin	biology
Kuwanna Dyer	geology
Leigh Justet	geology
Loyd West	geology
Mike Ginsberg	geology
Nicole Brown	geology
Placida Martinez	geology
Rebecca Kaufman	biology
Rebecca Kubart	geology
Robert Johnson	biology
Russell Turner	geology
Sylvain Brigant	computer science
Terri Waggoner	geology
Victoria Hansen	geology
Wendy Johnson	geology
Yrene Serna	computer science